

DOCUMENT CLASSIFYING DEVICE

Publication number: JP7114572

Publication date: 1995-05-02

Inventor: YUASA NATSUKI; UEDA TORU

Applicant: SHARP KK

Classification:

- international: G06F17/21; G06F17/27; G06F17/30; G06F17/21;
G06F17/27; G06F17/30; (IPC1-7): G06F17/30;
G06F17/27

- European:

Application number: JP19930259809 19931018

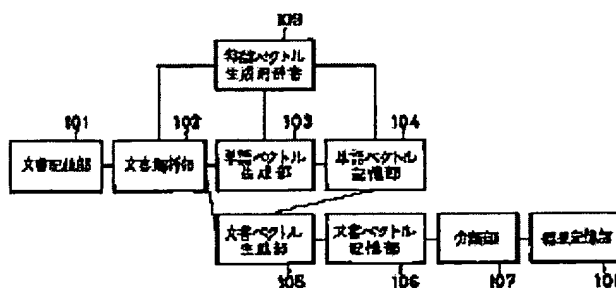
Priority number(s): JP19930259809 19931018

Report a data error here

Abstract of JP7114572

PURPOSE: To use semantic differences to automatically classify a document by automatically extracting feature vectors from the document and classifying the document based on these feature vectors.

CONSTITUTION: A storage part 101 where document data is stored, a document analysis part 102 which analyzes document data, a word vector generating part 103 which uses concurrent relations between words in the document to automatically generate a feature vector expressing the features of each word, a word vector storage part 104 where feature vectors are stored, a document vector generating part 105 which generates feature vectors of the document from feature vectors of words included in the document, a document vector storage part 106 where feature vectors of the document are stored, a classifying part 107 which uses the similarity between feature vectors of the document to classify the document, a result storage part 108 where the classification result is stored, and a feature vector generating dictionary 109 where words to be used for feature vector generation are registered are provided.



Data supplied from the esp@cenet database - Worldwide

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平7-114572

(43)公開日 平成7年(1995)5月2日

(51)IntCl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 17/30 17/27		9194-5L 7315-5L	G 0 6 F 15/ 401 15/ 20	3 1 0 D 5 5 0 F

審査請求 未請求 請求項の数3 O L (全 10 頁)

(21)出願番号 特願平5-259809

(22)出願日 平成5年(1993)10月18日

(71)出願人 000005049

シャープ株式会社

大阪府大阪市阿倍野区長池町22番22号

(72)発明者 湯浅 夏樹

大阪府大阪市阿倍野区長池町22番22号 シ
ャープ株式会社内

(72)発明者 上田 徹

大阪府大阪市阿倍野区長池町22番22号 シ
ャープ株式会社内

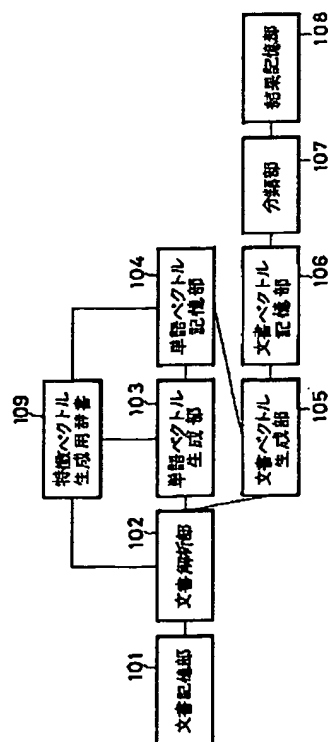
(74)代理人 弁理士 川口 義雄 (外1名)

(54)【発明の名称】 文書分類装置

(57)【要約】

【目的】 文書から自動的に単語の特徴ベクトルを抽出し、その特徴ベクトルをもとに文書を分類することで、意味的な異なりを用いた自動分類を可能にする。

【構成】 文書分類装置において、文書データを記憶する記憶部101と、文書データを解析する文書解析部102と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部103と、その特徴ベクトルを記憶する単語ベクトル記憶部104と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部105と、その特徴ベクトルを記憶する文書ベクトル記憶部106と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部107と、その分類した結果を記憶する結果記憶部108と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書109を備える。



【特許請求の範囲】

【請求項1】 文書分類装置において、文書データを記憶する記憶部と、文書データを解析する文書解析部と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部と、その特徴ベクトルを記憶する単語ベクトル記憶部と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部と、その特徴ベクトルを記憶する文書ベクトル記憶部と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部と、その分類した結果を記憶する結果記憶部と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書とを備え、大量の文書ファイル中の単語間の共起関係を用いて、各単語の特徴を表現する特徴ベクトルを自動的に生成し、文書を自動的に分類することができることを特徴とする文書分類装置。

【請求項2】 請求項1の文書分類装置の構成に加えて、結果記憶部に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部を備え、大量の文書ファイルを分類した後でその分類された各分類群ごとに単語の出現率を調べることで、分類に有用な単語を選出し、分類に有用な単語のみを用いることで分類の精度を向上させることができることを特徴とする文書分類装置。

【請求項3】 請求項1あるいは請求項2の文書分類装置の構成に加えて、結果記憶部に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める代表ベクトル生成部と、代表ベクトル生成部で生成された代表ベクトルを記憶する代表ベクトル記憶部とを備え、大量の文書ファイルを分類した後でその分類された各分類群ごとの単語や文書の特徴ベクトルを用いて、その分野を代表する特徴ベクトルを求めることができることを特徴とする文書分類装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、文書を保存／自動分類する文書自動分類機やワープロ／ファイリングシステムなどに利用される文書分類装置に関する。

【0002】

【従来の技術】 従来、文書の自動分類は困難であり、ユーザが手動で分類を行ったり、文書中のキーワードを抽出し、あらかじめ作成されたシソーラスを用いて分類を行っていた。また自動分類と称しているシステムでも分類のための基本的なデータは基本例文などの形で人手によって入力しておく必要があった。

【0003】

【発明が解決しようとする課題】 しかしながら、このような分類では人手による作業がボトルネックなるため、大量の文書の分類作業は大変困難である。

【0004】 本発明は以上の事情を考慮してなされたも

ので、人手を介することなく自動的に文書の分類を行なう文書分類装置を提供することを目的とする。

【0005】

【課題を解決するための手段】 請求項1に係る発明は、文書分類装置において、文書データを記憶する記憶部と、文書データを解析する文書解析部と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部と、その特徴ベクトルを記憶する単語ベクトル記憶部と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部と、その特徴ベクトルを記憶する文書ベクトル記憶部と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部と、その分類した結果を記憶する結果記憶部と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書とを備え、大量の文書ファイル中の単語間の共起関係を用いて、各単語の特徴を表現する特徴ベクトルを自動的に生成し、文書を自動的に分類することができることを特徴とする。

【0006】 また、請求項2に係る発明は、上記構成に加え、結果記憶部に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部を更に備え、大量の文書ファイルを分類した後でその分類された各分類群ごとに単語の出現率を調べることで、分類に有用な単語を選出し、分類に有用な単語のみを用いることで分類の精度を向上させることができることを特徴とする。

【0007】 また、請求項3に係る発明は、上記構成に加え、結果記憶部に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める代表ベクトル生成部と、代表ベクトル生成部で生成された代表ベクトルを記憶する代表ベクトル記憶部を更に備え、大量の文書ファイルを分類した後でその分類された各分類群ごとの単語や文書の特徴ベクトルを用いて、その分野を代表する特徴ベクトルを求めることができることを特徴とする。

【0008】

【作用】 請求項1での単語の特徴ベクトルの学習時の作用を説明する。文書記憶部に記憶されている大量の文書ファイルの内容が文書解析部に渡されて文の解析（形態素解析など）が行なわれ、単語ベクトル生成部で単語の共起関係や出現頻度等を分析して各単語の特徴ベクトルを生成する。こうして生成された単語の特徴ベクトルは単語ベクトル記憶部に記憶される。このようにして単語の特徴ベクトルの学習が行なわれる。特徴ベクトルを生成する単語は特徴ベクトル生成用辞書に登録されている単語に制限することで特徴ベクトルの記憶空間が巨大になりすぎるのを防ぐ。

【0009】 請求項1での文書の分類時の作用を説明する。文章の分類を行なう時には、文書記憶部に記憶され

ている文書ファイルの内容が文書解析部に渡されて文の解析（形態素解析など）が行なわれ、文書ベクトル生成部では文書解析部で文の解析をした時に出現する単語の特徴ベクトルを単語ベクトル記憶部を参照して求め、文書に含まれる単語の特徴ベクトルから文書の特徴ベクトルを生成する。こうして生成された文書の特徴ベクトルは文書ベクトル記憶部に記憶され、この文書の特徴ベクトル間の類似度によって分類部で文書を分類する。この分類結果は結果記憶部に記憶される。

【0010】請求項2に記載の構成では、大量の文書の分類を実行した後、有用単語選出部にて、結果記憶部に記憶されている分類結果を利用して分類時に有用な単語を選出する。有用単語選出部によって選出された単語だけを特徴ベクトル生成用辞書に登録してから再び単語の特徴ベクトルの学習を行なわせ、そうして得られた単語の特徴ベクトルを用いて再び分類を行なうことによって、請求項1の構成よりも特徴ベクトルの記憶空間を削減したり、また分類の精度をあげることもできる。

【0011】請求項3に記載の構成では、大量の文書の分類を実行した後、代表ベクトル生成部にて、結果記憶部に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める。代表ベクトル生成部で生成された代表ベクトルは代表ベクトル記憶部に記憶される。一度各分類群の代表ベクトルを生成してしまえば、新たな文書データを分類する時には、その文書の特徴ベクトルと各分類群の代表ベクトルとの比較を行なうだけでその文書がどの分類群に属するかを判定できる。

【0012】

【実施例】以下、本発明の好適な実施例を図面に基づき詳述する。

【0013】請求項1に係る発明の一実施例を図1に示す。文書分類装置は、文書データを記憶する記憶部101と、文書データを解析する文書解析部102と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部103と、その特徴ベクトルを記憶する単語ベクトル記憶部104と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部105と、その特徴ベクトルを記憶する文書ベクトル記憶部106と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部107と、その分類した結果を記憶する結果記憶部108と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書109とから構成される。

【0014】一般に通常の文書に使用されている単語の数は非常に多いため、特徴ベクトルを作成する際に用いる単語を制限しておく方が現実的である。このために用いるのが特徴ベクトル生成用辞書109で、ここに登録されている単語のみを用いて単語の特徴ベクトルを作成することで、特徴ベクトルの記憶空間の巨大化を抑える

ことができる。

【0015】図2は単語の特徴ベクトルの学習時のシステム構成を示し、単語の特徴ベクトルの学習時には、学習用の大量の文書データ文書記憶部101に記憶させておき、文書記憶部101から読み出した文書データは記事、段落、一文等の適当な塊ごとに文書解析部102に読み込まれ、文書解析部102でその文書データを解析して単語が抽出される。ここで抽出された単語列をもとにして単語ベクトル生成部103で単語の特徴ベクトルを生成し、単語ベクトル生成部103で生成された単語の特徴ベクトルは単語ベクトル記憶部104に記憶される。こうして単語の特徴ベクトルを学習する。

【0016】図3は文書分類時のシステム構成を示し、文書の分類をする時には、分類する文書のデータを文書記憶部101に記憶させておき、文書記憶部101から読み出した文書データは分類を行なわせたい単位（例えば記事単位）ごとに文書解析部102に読み込まれ、文書解析部102でその文書データの解析をして単語が抽出される。ここで抽出された単語の特徴ベクトルを104の単語ベクトル記憶部の内容を参照して求める。通常は文書データの一つの単位（例えば一つの記事）から複数の単語が抽出されるがこの時には求められるすべての単語の特徴ベクトルの値を平均化することで文書の特徴ベクトルが計算される。この時、単純に平均化するのではなく、各特徴ベクトルをその出現頻度の逆数に応じて重み付けをしてから（例えば、大量の記事からその単語の出現している記事数を調査し、 \log （全記事数/その単語が出現している記事数）をその単語の特徴ベクトルに掛けてから）平均化するとより良い値が得られる場合がある。

【0017】文書の特徴ベクトルが求まったら従来のクラスタリングの手法を適用することで文書の分類を行なうことができる。これは例えば文書の特徴ベクトル間の距離が近い文書同士は同じ分野に属するとみなせば良い。

【0018】また、人間が各分類群ごとに典型的な文書を選び、その文書から抽出される単語の特徴ベクトルからその分類群の仮の代表ベクトルを生成しておき、文書記憶部101から読み込まれる文書の特徴ベクトルがどの分類群の仮の代表ベクトルに近いかで文書を分類することもできる。このような分類手法でも文書記憶部101から大量に文書データを読み込ませれば仮の代表ベクトルを人間が選んでいるという誤差の影響が少なくなり、最終的には各分野毎のかなり一般的な代表ベクトルを生成することができる。

【0019】では具体的に単語の特徴ベクトルの生成法を説明する。単語の特徴ベクトルは、一塊の文書データの中に含まれている単語の出現頻度分布に、その単語のその一塊の文書データ中での出現頻度を掛けたものを加算していくことによって得られる。具体的な例で説明す

る。

【0020】例文A「アメリカ政府が先進主要国にココム規制の抜本的な見直しを提案してきた。」

例文B「規制対象国が兵器の製造につながる工業製品の輸出を規制することを条件に、ココムの規制品目を大幅に削減する意向のようだ。」

という文書データからどのように単語の特徴ベクトルを作成するかを説明する。ここでは、文書データは「一文」という単位で読み込まれることとするが、これは一記事など他の単位でも構わない。

【0021】また特徴ベクトルの次元数が21次元(特徴ベクトル生成用辞書に登録されている単語数が21)で各要素が「アメリカ、政府、先進、主要、国、ココム、規制、抜本的、見直し、提案、対象、兵器、製造、工業、製品、輸出、条件、品目、大幅、削減、意向」という単語に対応しているとする。

【0022】このような条件のもとで、例文Aが文書記憶部101から読み込まれると、文書解析部102が解析されて「アメリカ、政府、先進、主要、国、ココム、規制、抜本的、見直し、提案」が抽出される。この時単語ベクトル生成部103ではこれらの単語すべての特徴ベクトルのこれらの単語に対応する要素に1を加算する。すると、「アメリカ」「政府」等、例文Aに出現する単語の特徴ベクトルには(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)を加算する。これを図解したものが図8である。

【0023】次に例文Bが文書記憶部101から読み込まれると、文書解析部102で解析されて、「規制、対象、国、兵器、製造、工業、製品、輸出、規制、条件、ココム、規制、品目、大幅、削減、意向」が抽出される。

【0024】これから得られる単語出現頻度分布は(0, 0, 0, 0, 1, 1, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)である。「規制」は3回出現しているので、この単語出現頻度分布を3倍したベクトルである(0, 0, 0, 0, 3, 3, 9, 0, 0, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3)を「規制」の特徴ベクトルに加算し、「対象」「国」等、例文Bに1回しか出現しない単語の特徴ベクトルには(0, 0, 0, 0, 1, 1, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)を加算する。これを図解したものが図9である。

【0025】なお、このように常に整数を加算する方法では文の長さによって加算するベクトルの大きさが変化してしまうので、加算するベクトルの絶対値を1に正規化したり、出現頻度分布のベクトルの絶対値を1に正規化してから出現数に比例した値を掛けた後に加算する方法なども考えられる。

【0026】そして最終的に得られた特徴ベクトルは、

絶対値を1に正規化しておく。

【0027】こうして得られた単語の特徴ベクトルは単語ベクトル記憶部104に記憶され、文書の分類時に利用される。

【0028】次に、文書分類時の文書の特徴ベクトル生成の処理を、具体例として以下の例文Cが読み込まれた時をあげて説明する。

【0029】例文C「アメリカ政府は兵器の削減を提案した。」

10 例文Cが文書記憶部101から読み込まれると、文書解析部102で解析されて「アメリカ、政府、兵器、削減、提案」が抽出される。すると文書ベクトル生成部105では単語ベクトル記憶部104の内容を参照して「アメリカ」「政府」等、例文Cに出現する単語の特徴ベクトルを加算していき、例文Cの特徴ベクトルとして(3, 3, 3, 3, 5, 5, 9, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)を得る。これを図解したものが図10である。図10ではわかりやすさを優先するためにベクトルの正規化を行っていないが、実際の処理では加算する前に各単語の特徴ベクトルの絶対値を1に正規化してから加算を行なう。得られた特徴ベクトルは文書ベクトル記憶群106に記憶される。

20 【0030】次に、分類時に分類部107にて文書の特徴ベクトルをどのように利用するかを説明する。簡単には、まず求めた文書の特徴ベクトルの絶対値を1に正規化してから、K-means法などの従来からある手法を用いて分類したり、分類群の(仮)代表ベクトルとの類似度(距離を求めたり内積を計算することによって得られる)で分類すれば良いが、本手法で得られる特徴ベクトルは「多く出現する単語に対応する要素の値が非常に大きくなる」という特徴があるため、この特徴が分類に悪影響を与えないように工夫した方が良い分類結果が得られる場合が多い。例えば距離を求める場合には要素間の差が拡大しないような計算による距離(通常は各要素間の差の自乗和の平方根を計算するが、例えば各要素間の差の絶対値の和を計算するなどして求めた距離)を利用するようにしたほうが良いし、内積を求める前に全要素をlogをとったり冪乗根をとったりしてから正規化して値を均してから計算すると良い。

30 【0031】分類の具体例として、分類群が3つあり、それぞれの分類群の代表ベクトルが以下のように求められていたとしよう。

【0032】分類群1の代表ベクトル(1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

分類群2の代表ベクトル(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)

50 分類群3の代表ベクトル(4, 4, 4, 4, 6, 6,

7

6, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

類似度の尺度として、文書の特徴ベクトル、分類群の代表ベクトル共に絶対値を1に正規化してから両者の内積を計算し、一番大きな値をとるものが一番類似度が高いとすると、例文Cの特徴ベクトル

【0033】

【数1】

$$\frac{1}{\sqrt{238}}$$

【0034】(3, 3, 3, 3, 5, 5, 9, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)

分類群1の代表ベクトル

【0035】

【数2】

$$\frac{1}{\sqrt{8}}$$

【0036】(1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1)

分類群2の代表ベクトル

【0037】

【数3】

$$\frac{1}{\sqrt{285}}$$

【0038】(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)

分類群3の代表ベクトル

【0039】

【数4】

$$\frac{1}{\sqrt{210}}$$

【0040】(4, 4, 4, 4, 6, 6, 6, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

なので、例文Cの特徴ベクトルと各分類群の代表ベクトルとの内積は分類群1との内積

【0041】

【数5】

$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{8}} \cdot 20 \approx 0.4583$$

【0042】分類群2との内積

【0043】

【数6】

$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{285}} \cdot 150 \approx 0.5759$$

【0044】分類群3との内積

【0045】

【数7】

$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{210}} \cdot 211 \approx 0.9438$$

【0046】となり、例文Cの特徴ベクトルは分類群3

8

の代表ベクトルに一番近いことがわかるので、例文Cは分類群3に分類される。これを図解したのが図11である。図11も図10と同様わかりやすさを優先するためにベクトルの正規化を行っていないが、実際の処理では比較を行なう前に各ベクトルの絶対値を1に正規化してから比較を行なう。分類した結果は結果記憶部108に記憶される。

【0047】次に、本発明の請求項2の一実施例を図4に示す。ここで、符号201~209で表されるものは図1の符号101~109で表されるものと夫々同じものである。

【0048】文書分類装置は、文書データを記憶する記憶部201と、文書データを解析する文書解析部202と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部203と、その特徴ベクトルを記憶する単語ベクトル記憶部204と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部205と、その特徴ベクトルを記憶する文書ベクトル記憶部206と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部207と、その分類した結果を記憶する結果記憶部208と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書209と、結果記憶部208に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部210とからなる。

【0049】図5は学習時及び分類時のシステム構成を示す図である。最初は請求項1の実施例と同様の方法によって、単語の特徴ベクトルを学習し、それをもとに大量の文書データを分類する。分類した結果は結果記憶部208に記憶されているが、この結果を元にして、有用単語選出部210で有用単語の選出を行なう。これは、分類群ごとに各単語の頻度を求め、どの分類群にも同じような割合で含まれている単語を除去したり（方法1：最高頻度と最低頻度との比がある閾値以下のものを除去）、ある分類群にだけ高い割合で含まれているものを選出したり（方法2：最高頻度と第二位頻度との比がある閾値以上のものを選出）する。なお、有用単語選出部210で選出を行なう単語は必ずしも特徴ベクトル生成用辞書209に登録されている単語からでなくとも良く、もっと広い範囲の単語から選出を行なうことができる。

【0050】具体例として分類群がa, b, cの三つあったとして、特徴ベクトル生成用辞書209に登録されている単語が「政治、日本、国際」の三つだったとする。そして分類群ごとに各単語（特徴ベクトル生成用辞書209に登録されている単語以外に「選挙」「問題」についても頻度を調べるとする）の頻度が次のようだったとする。

【0051】分類群a 政治30%, 日本 5%, 国際 35%, 選挙10%, 問題20%

分類群b 政治 3%, 日本55%, 国際35%, 選挙2%, 問題 5%

分類群c 政治 3%, 日本30%, 国際35%, 選挙2%, 問題30%

すると、方法1を用いると「国際」はどの分類群にも同じような割合で含まれているので、特徴ベクトル生成用辞書から除去することになる。「政治」「日本」「選挙」「問題」は分類群ごとの頻度に偏りがあるので、有用単語として選出され、特徴ベクトル生成用辞書209に登録する(この時登録単語数を抑えたい場合は、頻度に偏りのある単語の中で、合計の出現頻度の順番に登録したい個数だけ取ってくれば良い)。方法2を用いた場合「政治」と「選挙」だけが選出され特徴ベクトル生成用辞書209に登録し、「日本」や「国際」や「問題」は特徴ベクトル生成用辞書209には登録しない。方法1と方法2の中間的な方法として、第1位の頻度と第n位(nは3以上、分類群の個数-1以下)の頻度との比がある閾値以上であるかどうかで有用単語を選出する方法も考えられる。また、頻度の比ではなく、頻度の分散の値が大きいものを選出する方法も考えられる。

【0052】なお、このようにして選出された単語は頻度の比(あるいは頻度の分散)に応じた重要度を持っていると考えることができるので、文書の特徴ベクトルを計算する時にはその文書内の単語の特徴ベクトルをこの比(あるいは分散)に応じて重み付けをしてから(例えば、 \log (頻度の比)をその特徴ベクトルに掛けてから)平均化するとより良い文書の特徴ベクトル地が得られる場合がある。

【0053】こうして特徴ベクトル生成用辞書209に、分類に有用な単語だけを登録し、もう一度、単語の特徴ベクトルを学習し、それを用いて文書を分類すると、特徴ベクトル生成用辞書をより小さくできたり、分類の精度をあげることができる。

【0054】本発明の請求項3の一実施例を図6に示す。ここで、符号301~310で表されるものは図4の201~210で表されるものと夫々同じものである。

【0055】文書分類装置は、文書データを記憶する記憶部301と、文書データを解析する文書解析部302と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部303と、その特徴ベクトルを記憶する単語ベクトル記憶部304と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部305と、その特徴ベクトルを記憶する文書ベクトル記憶部306と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部307と、その分類した結果を記憶する結果記憶部308と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書309と、結果記憶部308に記憶されている分

類結果を利用して分類時に有用な単語を選出する有用単語選出部310と、結果記憶部308に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める代表ベクトル生成部311と、代表ベクトル生成部311で生成された代表ベクトルを記憶する代表ベクトル記憶部312とからなる。

【0056】なお請求項1の実施例を用いて請求項3のシステムを構成する場合には有用単語選出部310が無いシステムとなる。

【0057】図7は学習時及び分類時のシステム構成を示す図である。最初は請求項1の実施例や請求項2の実施例と同様の方法によって、単語の特徴ベクトルを学習し、それをもとに大量の文書データを分類する。分類した結果は結果記憶部308に記憶されているが、この結果を元にして、代表ベクトル生成部311で代表ベクトルを生成する。これは例えば、分類群ごとの各単語の頻度を求め、ある分類群にだけ高い割合で含まれている単語を選出し、このような単語の特徴ベクトルの平均をとることによって生成できる。具体例として分類群がa, b, cの三つあったとして、特徴ベクトル生成用辞書309に登録されている単語が「政治、国会、国際」の三つだったとする。そして分類群ごとの各単語の頻度が次のようだったとする。

【0058】

分類群a 政治40%, 国会50%, 国際10%

分類群b 政治10%, 国会10%, 国際80%

分類群c 政治20%, 国会10%, 国際70%

すると、分類群aの代表ベクトルは、「政治」の特徴ベクトルと「国会」の特徴ベクトルの平均として与えられる。なお単なる平均ではなく、出現割合によって、重みをつけることも考えられる。例えば「政治」の出現頻度が「国会」の出現頻度の2倍なら、「政治」の特徴ベクトルの2倍と「国会」の特徴ベクトルとを加算し、3で割ったものを分類群aの代表ベクトルとする等である。

【0059】同様に分類群aに分類された文書の特徴ベクトルの平均をとったものを分類群aの代表ベクトルとする方法も考えられる。

【0060】こうして、代表ベクトルが生成されたらそれを代表ベクトル記憶部312に記憶しておくことで、以後の文書の分類時にはこの代表ベクトルを参照することで、文書記憶部301から読み込まれた文書は、その文書の特徴ベクトルにもっとも類似した代表ベクトルに対応する分類群に分類することができるようになる。

【0061】本発明は文書分類に用いるだけでなく、電子メールや電子ニュースを自動的に分類したり、電子メールの中や電子ニュースの中からユーザーの興味を持ちそうなものを選出したり(ユーザーがそれまでに読んだメールやニュースの特徴ベクトルとの類似度で判定できる)、あいまい検索(検索キーワードの特徴ベクトルと、検索対象文書の特徴ベクトルとの類似度が一定の閾

値以上になる文書を検索するようにすることで、検索キーワードに正確にマッチしていなくても関連のキーワードで検索できる)に利用できたり、仮名漢字変換における同音意義語の選択(それまでに変換した内容から得られる特徴ベクトルとの類似度で同音意義語を選択する)に利用できたり、音声認識・手書き文字認識などにおいて過去の文脈に最も適合した変換結果を選択する方法をとる(それまでに認識した内容から得られる特徴ベクトルとの類似度で認識結果を選択する)際にも利用できたり、認識時等において単語等の検索空間を狭める(それまでに認識した内容から得られる特徴ベクトルの要素のうち一定の閾値以上になっている要素に対応する単語だけを検索するようにする)際にも利用できる。

【0062】

【効果】本発明により、自動的に単語の特徴ベクトルを作成することができ、文書の分類を自動的に行なうことができるようになる。またこの方法で作成された単語の特徴ベクトルは文書の分類時だけでなく、あいまい検索や、仮名漢字変換における同音意義語の選択にも利用できるし、音声認識・手書き文字認識などにおいて、過去の文脈に最も適合した認識結果を選択する方法をとる際にも利用できる。

【図面の簡単な説明】

【図1】請求項1に係る発明の一実施例の基本構成を示すブロック図である。

【図2】図1に示すシステムの学習時のシステム構成を示すブロック図である。

【図3】図1に示すシステムの分類時のシステム構成を示すブロック図である。

【図4】請求項2に係る発明の一実施例の基本構成を示すブロック図である。

【図5】図4に示すシステムの学習、分類時のシステム構成を示すブロック図である。

【図6】請求項3に係る発明の一実施例の基本構成を示すブロック図である。

【図7】図6に示すシステムの学習、分類時のシステム構成を示すブロック図である。

【図8】単語の特徴ベクトルの生成を説明する図である。

【図9】単語の特徴ベクトルの生成を説明する図である。

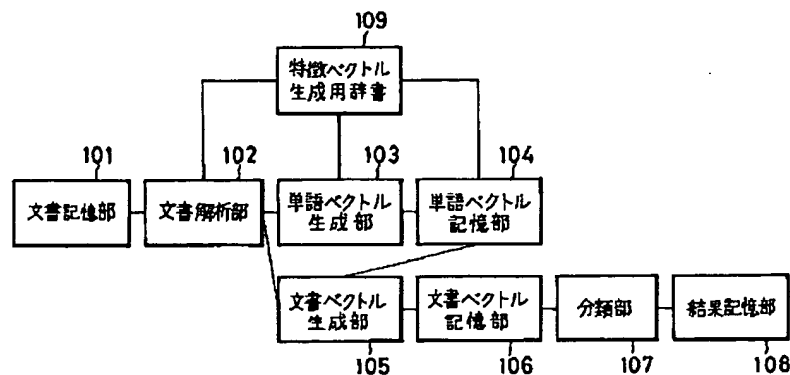
【図10】文書の特徴ベクトルの生成を説明する図である。

【図11】文書の分類を説明する図である。

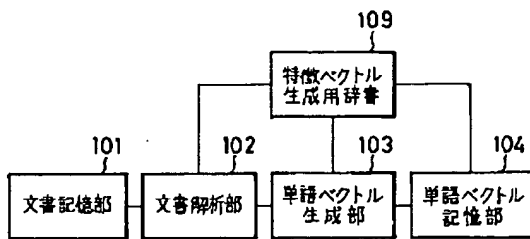
【符号の説明】

101、201、301 文書記憶部
102、202、302 文書解析部
103、203、303 単語ベクトル生成部
104、204、304 単語ベクトル記憶部
105、205、305 文書ベクトル生成部
106、206、306 文書ベクトル記憶部
107、207、308 分類部
108、208、308 結果記憶部
109、209、309 特徴ベクトル生成用辞書
210、310 有用単語選出部
311 代表ベクトル生成部
312 代表ベクトル記憶部

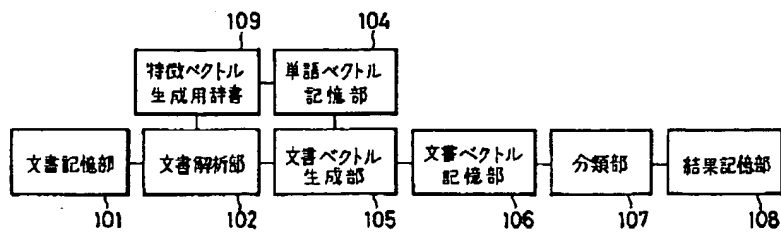
【図1】



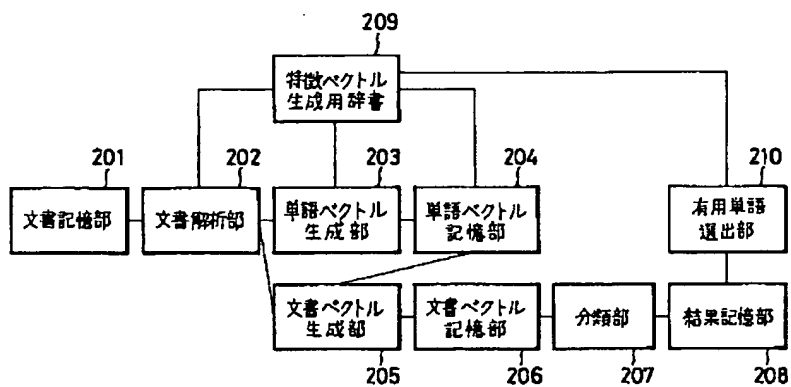
【図2】



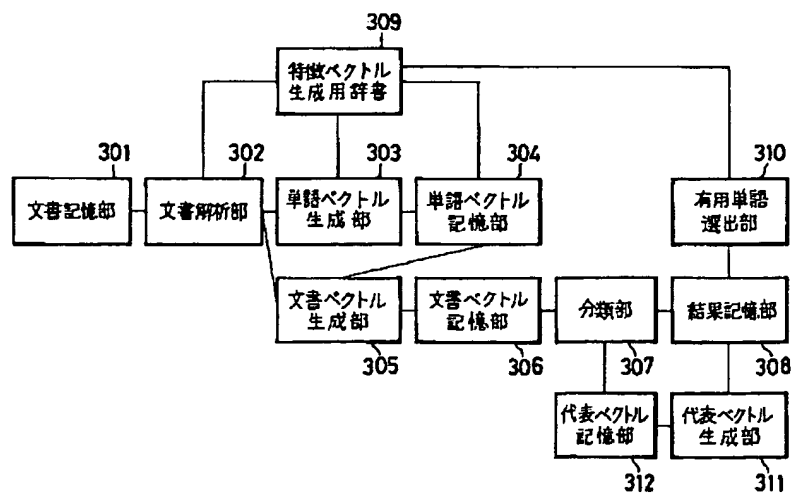
【図3】



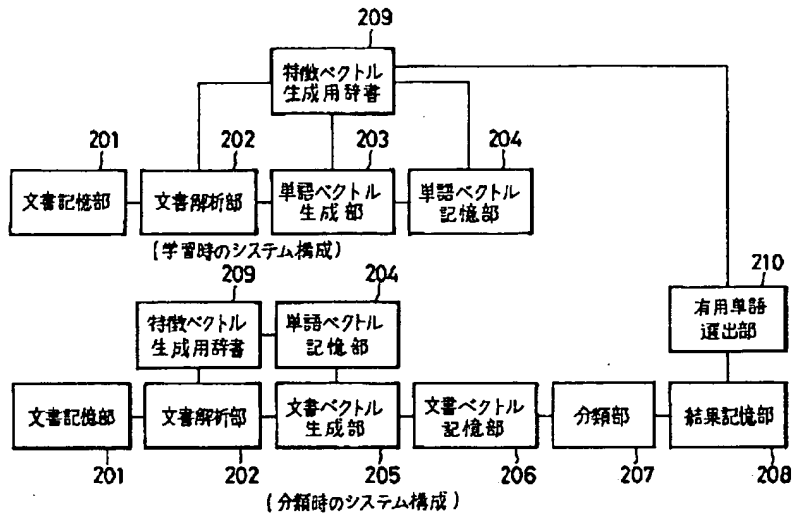
【図4】



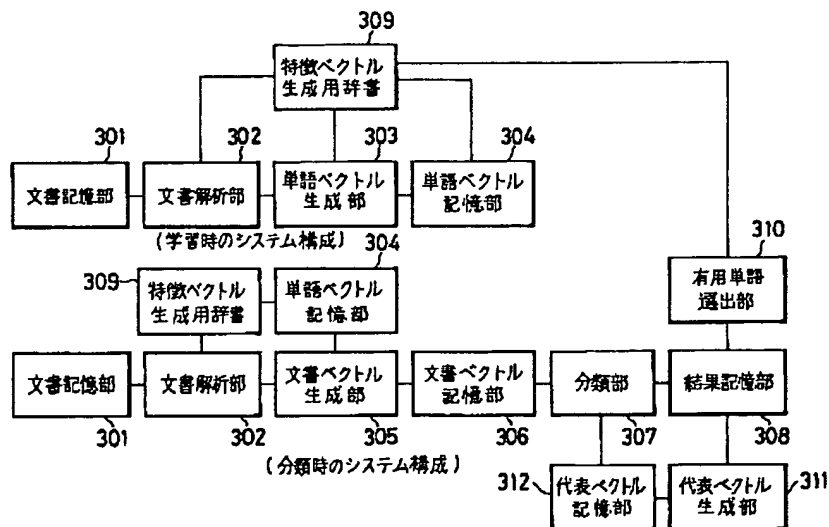
【図6】



【図5】



【図7】



【図10】

アメリカ (1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0)
 政府 (1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0)
 兵器 (0 0 0 0 1 1 3 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1)
 削減 (0 0 0 0 1 1 3 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1)
 +) 提案 (1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0)

例文Cの特徴ベクトル (3 3 3 3 5 5 9 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2)

【图 8】

[illegible]

【图 9】

[illegible]

【图 1 1】

例文Cの特徴ベクトル (3 3 3 3 5 5 9 3 3 3 2 2 2 2 2 2 2 2 2 2)

分類群1の代表ベクトル (1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1)

分類群2の代表ベクトル (1 1 1 1 1 1 1 1 1 1 5 5 5 5 5 5 5 5 5 5)

分類群3の代表ベクトル (4 4 4 4 6 6 6 3 3 3 1 1 1 1 1 1 1 1 1 1)

